Periodica Polytechnica Electrical Engineering and Computer Science

61(4), pp. 297-304, 2017

https://doi.org/10.3311/PPee.10414

RESEARCH ARTICLE

Monitoring Cognitive Workload Using Vocal Tract and Voice Source Features

Eydis Huld Magnusdottir[1*], Michal Borsky[1], Manuela Meier[1,2], Kamilla Johannsdottir[1], Jon Gudnason[1]

Received 15 December 2016; accepted 08 March 2017

Abstract

Monitoring cognitive workload from speech signals has received much attention from researchers in the past few years as it has the potential to improve performance and fidelity in human decision making. The bulk of the research has focused on classifying speech from talkers participating in cognitive workload experiments using simple reading tasks, memory span tests and the Stroop test, typically into three levels of low, medium and high cognitive workload. This study focuses on using parameters extracted from the vocal tract and the voice source components of the speech signal for cognitive workload monitoring. The experiment used in this study contains 98 participants, the levels were obtained by using a reading task and three Stroop tasks which were randomly ordered for each participant and an adequate rest time was used inbetween tasks to mitigate the effect of cognitive workload from one task affecting the subsequent one. Vocal tract features were obtained from the first three formants and voice source features were extracted using signal analysis on the inverse filtered speech signal. The results show that on their own, the vocal tract features outperform the voice source features. The MCR of 33.92% ± 1.05 was achieved with a SVM classifier. A weighted combination of vocal tract and voice source features classified with SWM classifier fused at the output level achieved the lowest MCR of 32.5%.

Keywords

speech science, voice source signal, vocal tract features, computational paralinguistic

[1] Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland

[1] Department of Electrical and Computer Engineering, Technical University of Munich, Germany

[*] Corresponding author, e-mail: eydis07@ru.is

1 Introduction

Monitoring cognitive workload in individuals that are performing safety-critical jobs has huge potential, for example in aviation [1]. The relationship between cognitive workload and performance has been well studied [2] and the connection between cognitive workload and physical health has also been highlighted [3]. Clearly, it is crucial to manage cognitive workload in the modern day work environment, both in terms of performance and health. Speech processing offers the ability to monitor cognitive workload in a non-intrusive way, compared to measures such as blood pressure, heart rate, electroencephalogram or electrocardiogram. Measuring voice is easy and recent work has shown very promising results in classifying cognitive workload levels using the speech signal [4-6]. Successful design and implementation of such a method would provide a powerful tool to developers of cognitive infocommunications systems [7].

The main objective of this work is to provide an independent verification of whether there is a link between increased cognitive workload and changes in the speech signal and, if this link exists, to characterize what part of voice is mostly affected. To do this we conducted a cognitive workload experiment where the set tasks were primarily solved using speech from 98 participants. Each participant read a standard passage of text on a computer screen and solved Stroop tasks with three difficulty levels [8]. A separate classifier of the three difficulty levels was trained for each speaker based on two sets of voice parameters comprising of the vocal tract (VT) and voice source (VS) features respectively. The conclusion is that task load does affect the speech signal and that vocal tract parameters are a better indicator of task level than voice source parameters. The context of the work is presented in Section 2 and the speech processing approach is described in Section 3. The experimental methodology and detailed results are given in Section 4 and 5. The results are summarized in Section 6 and the discussion is continued in Section 7.

2 Speech processing for cognitive workload

There is a growing body of work that supports the statement that cognitive workload affects the speech signal. For example, a data set of 15 participants was presented in [4] where each participant performed reading and Stroop tasks. The difficulty levels of these tasks were classified using mel-frequency cepstrum coefficients and prosodic features and a classifier based on a speaker-adapted Gaussian mixture model. The feature extraction was extended to include targeted extraction of vocal tract features through sub-band centroids [5]. These classification schemes indicated a strong relationship between cognitive workload and the speech signal within the experimental framework of the studies. The INTERSPEECH 2014 Cognitive Load Challenge (ComParE) was based on this methodology. A data set contained 26 participants providing speech recordings and an electroglottograph during a reading task and a low-, medium-, and high cognitive load level Stroop tasks [9]. The winning entry used an i-vector classification scheme on a combined feature set of fused speech streams, prosody and phone-rate [6]. Other entries used, for example, sub-band centroid features [10] and voice quality features [11] supporting earlier findings.

Other work that does not rely on induced cognitive load levels via tasks includes experiments done in military flight simulation [12]. Mean change in fundamental frequency and speech intensity was shown to increase and their range decreased as cognitive load of the flights increased. Urban search and rescue training operations were recorded and annotated for cognitive load analysis [13] using prosodic features, Teager energy, voicing strengths and spectral envelope.

This work focuses on two sets of voice features extracted from the vocal tract and the voice source respectively. Voice can be decomposed into the vocal tract and the voice source using the traditional linear source-tract model [14, 15]. The vocal tract can be modeled using an auto-regressive moving average model corresponding to the formants and anti-formants. These were extracted using extended Kalman smoothing [16] and used as vocal tract features for analyzing depression [17].

Voice source features can be extracted indirectly from the speech signal via covariance analysis and cepstrum processing without relying on inverse filtering [18,19]. Various methods for inverse filtering have however been implemented and evaluated [20] and feature extraction methods based on the glottal flow estimate have been proposed [21, 22, 23]. A set of voice source features was studied in relation to short affect bursts produced by 10 professional actors [24].

In this work vocal tract features are extracted using the Kalman smoothing approach [16] and the features described in [24] were used for the voice source. The two feature sets were combined using the temporal cross-correlation structure described in [17], which has been proven to be effective for detecting emotions and depression in speech and other signals. The cross-correlation structure simultaneously captures the short- and long term relationship between individual features (vocal tract and voice source) in the feature vector. The detail is described in the following section.

3 Voice Feature Extraction

The approach adopted in this work was based on the idea that decomposing the speech signal into the vocal tract characteristics and the voice source signal gives a better insight into the voice changes brought by increased cognitive workload. Two sets of distinct features were extracted from the speech signal, the first set consisted of features describing the vocal tract and the second set consisted of features describing the voice source signal.

3.1 Vocal Tract Features

The set of vocal tract features used in this work consisted of ordinary formants F1, F2 and F3. These were extracted from the speech signal using the KARMA [16] algorithm. It works on the principle of Kalman-based auto-regressive moving average smoothing. The main advantage of using KARMA is that the algorithm produces smoother formant contours than other methods and they have less erratic spikes around voiced-unvoiced transitions. The authors have demonstrated that this approach exhibits lower overall root-mean-square error and thus can be considered as more reliable. The three formants were obtained for each 20 ms of speech and concatenated in a vocal tract feature vector $\mathbf{x}_{vt}(j)$, where j denotes the frame index. Fig. 1b illustrates an output of KARMA algorithm for a given speech signal on Fig. 1a.

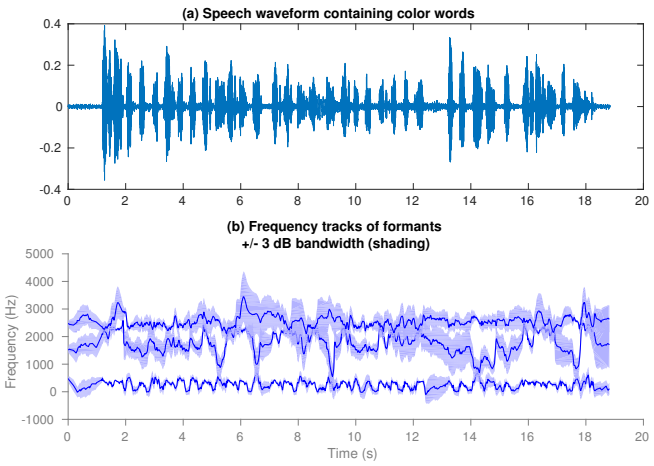

Fig. 1 The result of formant tracking using KARMA algorithm.

3.2 Voice Source Features

The voice source features consisted of 10 different parameters that were extracted either directly from the speech signal or from an estimate of the glottal flow or its derivative (voice source signal). The following is a list of the parameters and their time- or frequency domain association:

- time - maximum flow declination rate (MFDR), pulse amplitude (PA), normalized amplitude quotient (NAQ) and closed quotient (CQ),
- freq. - fundamental frequency (f0), harmonics to noise ratio (HNR), harmonics richness factor (HRF), 1^{st} to 2^{nd} harmonics ratio (H1-H2), jitter (Jitt) and cepstral peak prominence (CPP).

Cepstral peak prominence (CPP) was the only measure extracted directly from the speech signal and has been historically widely used to classify and rate levels of dysphonia [25]. While there is no clear understanding of what the parameter measures, the general findings among the speech pathologists show that the parameter is tied to vocal attributes of breathiness, roughness and hoarseness. A theoretical study on the parameter's nature was done in [26] where the authors concluded that CPP integrates measure of several features describing the aperiodicity and waveform of the acoustic voice signal.

The iterative adaptive inverse filtering algorithm [27] was employed in order to obtain the voice source signal from which the rest of the voice source measures were extracted. Authors in [24] studied the relation of these features to the emotions and found a statistical significance for all of them aside from NAQ. However, other works [28] have reported on increase of this parameter for hypo-functioning (e.g. relived) voice and a decrease for hyper-functioning voice (e.g. angered) [29].

The voice source features were obtained for each 20 ms of speech and concatenated in a voice source feature vector $\mathbf{x}_{vs}(j)$, where j denotes the frame index.

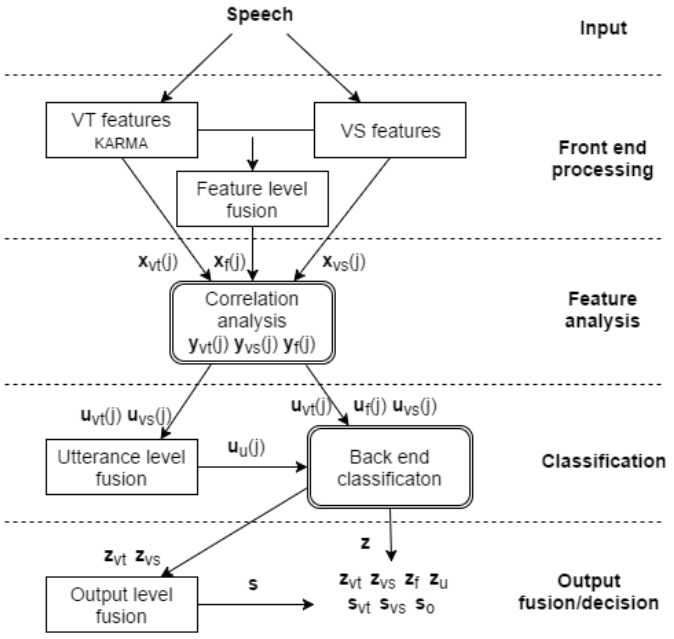

Fig. 2 Fusion levels and processing flow block scheme

3.3 Feature level fusion

Feature level fusion was achieved by concatenating the frame-level feature vectors of the vocal tract features ($n_{c-vt} = 3$ formant values) and voice source features ($n_{c-vs} = 10$ parameters) for each frame, so that, instead of having only $\mathbf{x}_{vt}(j)$ or $\mathbf{x}_{vs}(j)$, we now have $\mathbf{x}_f(j) = [\mathbf{x}_{vt}(j)^T, \mathbf{x}_{vs}(j)^T]^T$ for each frame j. From this a correlation matrix (with $n_c = 13$ parameters) and an utterance-level feature vector $\mathbf{u}$ was made. Fig. 2 depicts the steps included in the processing flow and at which level the different fusion schemes took place.

3.4 Temporal correlation structure

The three feature streams $\mathbf{x}_{vt}(j)$, $\mathbf{x}_{vs}(j)$ and $\mathbf{x}_f(j)$ are processed further and summarized to produce a single feature vector $\mathbf{u}_{vt}$, $\mathbf{u}_{vs}$ or $\mathbf{u}_f$ for the utterance to be classified. This is done by using the correlation structure of the feature stream [17]. A fixed time scale of 2 frames was used to create a concatenated feature vector of the current feature vector at time j and 13 successive time delays. For the joint vocal tract and voice source vector $\mathbf{x}_f(j)$ the new data vector of $N = n_c n_d = 13 \times 14 = 182$ dimensions is created with,

$$\mathbf{y}_f(j) = \left[\mathbf{x}_f^T(j), \mathbf{x}_f^T(j-2), \mathbf{x}_f^T(j-4), \ldots \mathbf{x}_f^T(j-26)\right]^T. \tag{1}$$

The vocal tract $\mathbf{y}_{vt}(j)$ and voice source feature vectors $\mathbf{y}_{vs}(j)$ therefore have dimensions $N_{vt} = n_{c-vt} n_d = 13 \times 14 = 42$ and $N_{vs} = n_{c-vs} n_d = 10 \times 14 = 140$ dimensions respectively.

The cross-correlation matrix for the utterance is then formed using,

$$\mathbf{R} = \frac{1}{N_s} \sum_j \tilde{\mathbf{y}}(j) \tilde{\mathbf{y}}^T(j) \tag{2}$$

where N_s is the number of frames in a utterance and $\mathbf{y}(j)$ is normalized to have a zero mean and unit variance, denoted as $\tilde{\mathbf{y}}(j)$. The eigenvalues λ_i are obtained from $\mathbf{R}$ in a descending order and each component in the utterance level feature vector $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T$ is computed as the normalized cumulative sum

$$u_n = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{3}$$

where $n = 1,2,\ldots, N$. The component u_n therefore represents the proportion of energy contained in the first n eigenvectors.

Fig. 3 illustrates the extracted vocal tract $\mathbf{u}_{vt}$ and voice source $\mathbf{u}_{vs}$ feature vectors for three cognitive workload classes (explained in next section). The figure demonstrates that for these utterances for the vocal tract features, more than 90% of the energy is captured by the first 8 out of $N_{vt} = 42$ components whereas for the voice source features more than 43 out of $N_{vs} = 140$ components are needed to capture 90% of the energy. The number of eigenvalue components in the figure are

truncated for illustrative purposes after the cumulated eigenvalues reached 99.9%.

3.5 Feature level fusion

Feature level fusion was achieved by concatenating the frame-level feature vectors of the vocal tract features ($n_{c-vt} = 3$ formant values) and voice source features ($n_{c-vs} = 10$ parameters) for each frame, so that, instead of having only $\mathbf{x}_{vt}(j)$ or $\mathbf{x}_{vs}(j)$, we now have $\mathbf{x}_f(j) = [\mathbf{x}_{vt}(j)^T, \mathbf{x}_{vs}(j)^T]^T$ for each frame j. From this a correlation matrix (with $n_c = 13$ parameters) and an utterance-level feature vector $\mathbf{u}$ was made. Fig. 2 depicts the steps included in the processing flow and at which level the different fusion schemes took place.

3.6 Utterance level fusion

Utterance level fusion was achieved by concatenating the high-level feature vectors $\mathbf{u}_{vt}$ and $\mathbf{u}_{vs}$ into a single vector $\mathbf{u}_u$. This approach is distinctly different from the frame-level fusion. Feature-level fusion assumes that modeling cross-correlations between vocal tract and voice parameters might provide new information about the cognitive workload. Utterance-level fusion, on the other hand, assumes that these two streams are independent of each other and thus can be used in tandem.

3.7 Output level fusion

The output level fusion was performed by taking the soft scores from a classifier for both vocal tract $\mathbf{z}_{vt}$ and voice source $\mathbf{z}_{vs}$ streams and accumulating them into a single score. The weights for particular streams were subjected to sensitivity analysis on the test set in order to determine their optimal values and had to meet a simple criteria $\sum_i w_i = 1$. The accumulated scores were then classified using the same maximum a posteriori criteria.

4 Methodology

4.1 Experiment

Our experiments were designed around the Stroop test [8], which shows words for colours, such as 'blue' and 'red' on a computer screen in either congruent colour (the word 'red' is shown in red), or in-congruent colour (the word 'red' is shown in, for example, blue). Participants are instructed to name the colours in which the words are shown as quickly as possible. If the word is shown in in-congruent colour, the task is likely more difficult than when it is shown in congruent colour as the brain is wired to automatically read words when shown, pre-loading the brain with the wrong information (colour name). An additional increase in task difficulty was created by adding a rather strict time limit for naming the colours and only showing a single coloured word on the screen at once, preventing participants from 'looking ahead'. We used the same five colours and corresponding words across the three conditions and the words were shown in the native language of the participants (Icelandic). All speech was recorded using a head-mounted microphone and the speech was sampled at 48 kHz.

A total of 98 participants visited the laboratory over a period of three months, 27 identified as male and 71 as female. Further statistics of the participants are summarized in Table 1.

Table 1 Age, height and weight of the participants in the study.

	Mean	St.dev.	Min.	Max
Age [years]	25.2	5.73	18	53
Height [cm]	172.4	7.95	153	193
Weight [kg]	73.5	13.3	50	114

The participants started with a 10 minute resting period during which baseline values for their cardiovascular activity was determined. Ideally, the subjects would be completely at rest. Then all subjects were asked to read aloud a short passage of text which took around 2 minutes. It aims to engage the subjects in the same physical activity - speaking - as during the Stroop tasks, but performing a task (reading) that they were likely experienced in and comfortable with. After the reading task followed 3 minutes of recovery time (rest), before starting on the Stroop tests. Three Stroop difficulty levels, congruent (L1), mostly in-congruent (L2), and time-limited (L3), were presented in random order (Latin square with 6 possible orders) and with 5 minute intervals between each Stroop level. Each set consisted of 6 screens with 35+1 words each (216 words per Stroop level), where the last word was used to indicate the end of the screen. The time-limited task had an randomly alternating limit of 0.75 or 0.65 seconds per word, which was chosen for being roughly the amount of time participants needed to do the non-time-limited tasks (0.75s) and a bit less so as to ensure the task would be considered extra hard (0.65s). The average time taken to complete a single screen of the congruent task was 23 seconds, increasing to 30 seconds for the in-congruent task, and 29 and 25 seconds for the time-limited tasks. Each screen was recorded as a single utterance from which an utterance level feature vector **u** is extracted as explained in Section 3.

4.2 Classifiers & Evaluation

The primary goal of this paper was to answer the questions: "*Can increased workload be detected in the recordings from a series of Stroop tests? And if yes, what features are best suited for this task?* Thus, the experimental task was to construct a trinary classifier that would classify unknown recordings into three different Stroop levels. In order to measure the discriminatory capabilities of the discussed features without introducing additional speaker variability, the classifiers were built as speaker-dependent.

Three distinct types of supervised classifiers were used in this study: Minimum distance (MD) to the class centroid,

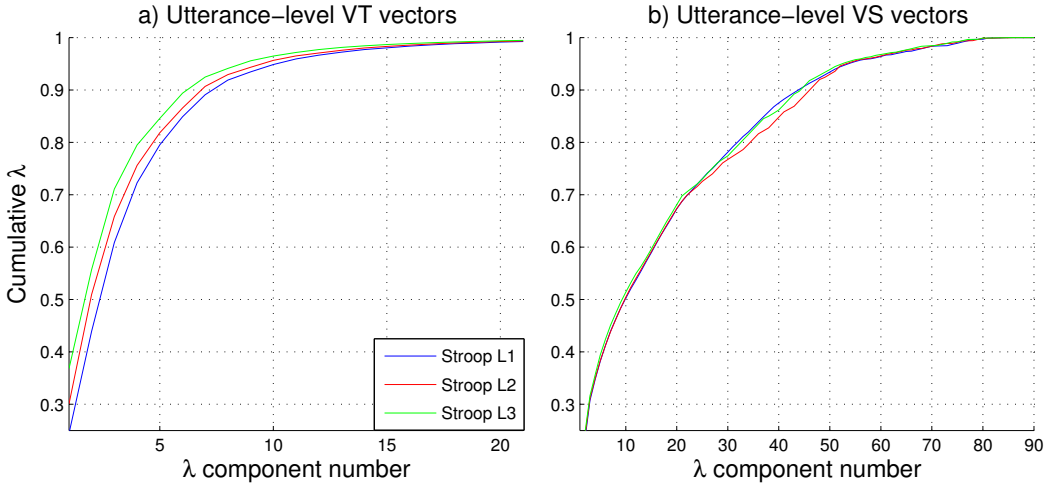

Fig. 3 An example of a vocal tract and voice source features for all three Stroop tasks

Support Vector Machine (SVM) and Random Forests (RF). The MD based classifier represents the most intuitive approach to the classification which makes it an ideal choice to gain an insight into the data's separability using chosen features. A class centroid was calculated from the training data and the classification was based on the minimum Euclidean distance between the centroids and the test vector. While the SVM is fundamentally a binary classifier, it is also possible to combine multiple SVMs and solve multiclass classification problems. The approach taken in this article was based on constructing all possible two-class SVMs and leaving the MD to resolve the ambiguous cases. The SVMs used linear kernel function, the soft margin optimization method, the hyperparameter C was set to 1 and no observations were allowed to violate the Karush-Kuhn-Tucker conditions. The third classifier was based on random forests, when the final ensemble had 20 trees and the minimum number of observations in a leaf was set to 1.

The amount of available data did not allow us to draw statistically valid conclusions from just a single split of the dataset into training and testing sets. As a consequence, we used a leave-one-out cross validation strategy. The testing set consisted of a single signal, while the rest of the signals (20) were put into the training set to train the classifier. And finally, the evaluation results was recorded into a cumulative confusion matrix. This whole process was repeated for each signal separately, which allowed us to train 21 classifiers and to obtain 21 classification scores for every possible set division. The final confusion matrix after the whole run was used to compute the total misclassification rate (MCR) that is presented in the tables below.

4.3 Output level fusion

In the final approach the fusion was done on the output of the classifiers. As the results show in Table 3, out of the three classifiers the SVM outperforms them in all instances. As a result the output level fusion was performed only for the the SVM classifier.

A cross confusion table comparing the classification results of the SVM for vocal tract $\mathbf{z}_{vt}$ and voice source $\mathbf{z}_{vs}$ feature streams gives the indication of this fusion being beneficial. Table 2 shows that the MCR for both VS and VT classifying correctly to be 38.11%. The results in the off diagonal line of the table show that MCR for VS classifying correctly and VT incorrectly is 15.87% and 27.40% for VS incorrect and VT correct. These are the results that might be influenced to give even better results for combined classification.

Table 2 Cross confusion table for classification performance [%] between vocal tract features and voice source features with the SVM classifier.

	VS Correct	**VS Incorrect**
VT Correct	**38.11**	27.40
VT Incorrect	15.87	18.62

The actual process of converting the distance metric, which is the standard output of the SVM classifier, into soft scores was based on the following hypothesis. The further the point is from a margin, the greater the confidence it belongs to that particular class. As a consequence, the values for both streams were simply added together with the corresponding weights and classified, with one deviation. Instead of leaving MD to resolve the ambiguous cases, the class with the highest score was chosen. A sensitivity analysis was performed to get insight into the optimum combination of weights w between $\mathbf{s}_{vs}$ and $\mathbf{s}_{vt}$ feature streams.

$$\mathbf{s}_o = w\mathbf{s}_{vs} + (1-w)\mathbf{s}_{vt} \tag{4}$$

The results of this is presented in Fig. 4 where a optimum combination of weights between $\mathbf{s}_{vt}$ and $\mathbf{s}_{vs}$ seems to occur in the interval of weights $w = 0.2 - 0.4$.

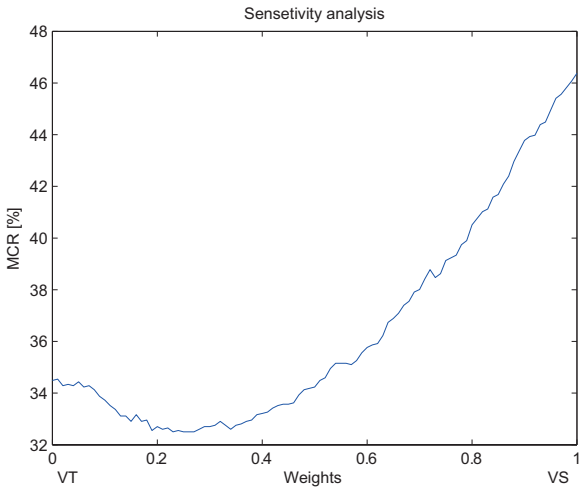

Fig. 4 Sensitivity analysis on MCR[%] between combined vocal tract $\mathbf{s}_{vt}$ and voice source $\mathbf{s}_{vs}$ classification results.

5 Results

The results obtained from the classification tests are shown in Table 3. There are several interesting things that can be highlighted. The first thing to look at is the performance of the proposed speech features in the task of classifying the Stroop level and its correlation to the cognitive workload. A trinary classification task with a set of completely random features would theoretically achieve the MCR of 66%, but the results for our best features were below 33%. These results indicate that the cognitive workload causes changes to the voice that can be observed and objectively measured.

Before the output level fusion was applied the vocal tract features achieved better MCRs than any other set of features. The overall best results of 33.92% were obtained with a SVM classifier, while the RF and MD classifiers followed with the MCRs of 43.15% and 43.93% respectively. The second best results were achieved with combined features (both utterance and feature-level fused) and the voice source features scored as the last. Another interesting thing to note is the fact that utterance-level fusion outperformed the feature-level fusion by 5.49% for SVMs. In all studied setups, the SVMs proved to consistently outperform other classifiers.

Table 3 Average MCR [%] over all speakers with different sets of features and classifiers.

Features	MD	SVM	RF
VTfeat $\mathbf{z}_{vt}$	43.93±1.09	**33.92±1.05**	43.15±1.09
VSfeat $\mathbf{z}_{vs}$	65.65±1.05	47.47±1.1	55.59±1.1
Utt. Fused $\mathbf{z}_{u}$	53.43±1.08	35.86±1.06	41.77±1.09
Feat. Fused $\mathbf{z}_{f}$	42.91±1.09	41.35±1.09	49.08±1.1

The output level fusion method turned out to achieve the lowest MCR for all the methods. The combined vocal tract and voice source with the weight value of $w = 0.24$ turned out to have 32.5% MCR. A comparison of the output level fusion results is presented in Table 4. In the table the results of the SVM classification scheme and the soft score classification are represented. The results show that by applying the output level fusion method the combined feature streams, vocal tract and vocal source, outperform the result of vocal tract feature stream and SVM classification combination already presented.

Table 4 Comparison of soft score classification and SVM classification. MCR [%] between vocal tract features and voice source features.

	Soft score s	SVM z
VT feat	34.49	33.92
VS feat	46.38	47.47
Utt.fusion	-	35.86
Output fusion	**32.50**	-

An insight into the separability of the distinct Stroop levels can be achieved by taking a closer look at a confusion matrix. Table 5 presents the results for SVM classifiers with the vocal tract features averaged over all speakers. The actual Stroop L1 utterances were more often confused for Stroop L2 tasks then they were for the L3 tasks. The same trend, although naturally reversed, was observed for actual Stroop L3 utterances. This trend was much more pronounced for actual Stroop L3 tasks. This observation leads us to the conclusion that an increase in the cognitive workload increases the changes in the voice.

Another interesting fact is that a classifier trained for Stroop L2 tasks misclassified the Stroop L3 features more often than Stroop L1. This observation leads to the conclusion that the increased cognitive difficulty of Stroop L2 and Stroop L3 tasks introduces changes to the voice, which can be accurately detected using the vocal tract features.

Table 5 MCR [%] matrix for the vocal tract features with the SVM classifier.

	Stroop L1	Stroop L2	Stroop L3
Stroop L1	**72.5**	15.7	11.8
Stroop L2	17.5	**59.7**	22.8
Stroop L3	10.1	24.6	**65.3**

A histogram of MCR for all 98 participants is presented in Fig. 5 highlighting several interesting aspects concerning the individuals performance in the tasks and its reflection on the classification results. A single participant scored MCR of 60%, which is very close to random classification, while a single participant scored MCR of 0%, which represents a perfect classification score. Never the less, most of the participants scored around the average MCR of 33%.

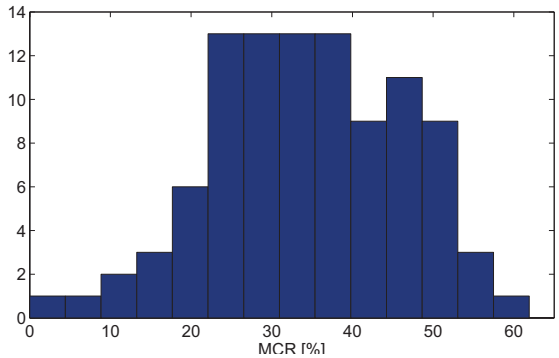

Fig. 5 A histogram of MCR [%] for the whole set of speakers.

6 Conclusions

This study gives an independent verification of the relationship between cognitive workload and speech. Speaker independent classifiers trained on 98 participants were able to distinguish between utterances of low, medium and high cognitive workload with 32.5% misclassification rate with the combination of vocal tract and vocal source features fused at the output level. The vocal tract features achieved 33.92% misclassification rate and the voice source features only achieved 46.38% misclassification rate on their own.

Specific conclusions of this work are that the particular vocal tract parameters used in this work [16] outperform the voice source parameters that are studied in [24] for emotion detection of which some are also used for cognitive workload classification in [22]. This is in concurrence with studies that compare vocal tract and voice source features both for other tasks [18, 19] and cognitive workload [22]. Never the less the performance of the vocal tract parameters can be improved by fusing the two parameters at the output level.

More generally, the study reinforces previous findings that show that there is a link between cognitive workload and speech [12, 22, 11, 6, 13].

7 Discussion

Speaking is a cognitive activity so it is not surprising that a simultaneous task affects the voice. Misclassification rate of over 30% is high however, given that each test utterance is over 20 s in duration and the number of classes is only three. There are two possible reasons behind this discrepancy. The first is related to precisely that of speaker identification but individuals react differently to cognitive workload [2]. This study has tried to overcome this by training individual classifiers for each participant and other studies do similar things, for example using speaker adaptation [22]. Better methods for addressing individual differences for cognitive workload are therefore called for. The second reason that explains the classification performance has to do with the dynamic nature of the body's response to cognitive workload [30]. Evidence from studies using cardiovascular reactivity shows that the body's response to cognitive stimulus is strong during the first few seconds of the task but fades as the participant gets used to the task. This dynamic behavior is not captured in the current experimental framework used for speech classification but there is a strong reason to suggest that a time-varying analysis would improve classification.

In another sense, the results in this paper give reasons to optimism. The results show the voice responds differently to different task load levels. The data gathering setup gives us an opportunity to adapt the methodology to the concerns regarding the time-varying nature of the body's response to cognitive workload. This is indeed the focus of our future research.

Acknowledgement

The work was performed during Manuela Meier's Erasmus+ traineeship at Reykjavik University. This work is sponsored by The Icelandic Centre for Research (RANNIS) under the project Monitoring cognitive workload in ATC using speech analysis, Grant No 130749051.

References

[1] Experimental Centre Note No 18/06. "Evaluation of the human voice for indications of workload induced stress in the aviation environment." Tech. Rep., 2006

[2] Galy, E., Cariou, M., Mélan, C. "What is th erelation ship between mental work load factors and cognitive load types?." *International Journal of Psychophysiology*. 83(3), pp. 269–275. 2012. https://doi.org/10.1016/j.ijpsycho.2011.09.023

[3] Kompier, M. A., Aust, B., van den Berg, A. M., Siegrist, J. "Stress prevention in bus drivers: evaluation of 13 natural experiments." *Journal of Occupational Health Psychology*. 5(1), pp. 11–31. 2000.

[4] Yin, B., Chen, F., Ruiz, N., Ambikairajah, E. "Speech-based cognitive load monitoring system." In: *2*008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 31 March - 4 April 2008, pp. 2041-2044. https://doi.org/10.1109/ICASSP.2008.4518041

[5] Le, P. N., Ambikairajah, E., Epps, J., Sethu, V., Choi, E. H. "Investigation of spectral centroid features for cognitive load classification." *Speech Communication*. 53(4), pp. 540–551. 2011. https://doi.org/10.1016/j.specom.2011.01.005

[6] Van Segbroeck, M., Travadi, R., Vaz, C., Kim, J., Black, M. P., Potamianos, A., Narayanan, S. S. "Classification of cognitive load from speech using an i-vector framework." In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, Sept. 14-18, 2014.

[7] Baranyi, P., Csapo, A. "Cognitive infocommunications: coginfocom." In: *1*1th IEEE International Symposium on Computational Intelligence and Informatics CINTI 2010: Proceedings. (Szakal, A. (ed.)). pp. 141–146. Budapest, Hungary, Nov. 18-20, 2010, pp. 141-146. https://doi.org/10.1109/CINTI.2010.5672257

[8] Stroop, J. R. "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology.* 18(6), pp. 643–662, 1935. https://doi.org/10.1037/h0054651

[9] Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y. "The interspeech 2014 computational paralinguistics challenge: cognitive & physical load." In: INTERSPEECH 2014. pp. 427–431. 2014.

[10] Kua, J. M. K. Sethu, V., Le, P. N., Ambikairajah, E. "The unsw submission to interspeech 2014 compare cognitive load challenge." In: INTERSPEECH 2014, pp. 746–750.

[11] Huckvale, M. A. "Prediction of cognitive load from speech with the VOQAL voice quality toolbox for the interspeech 2014 computational paralinguistics challenge." In: InterSpeech 2014, Singapore.

[12] Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., Leino, T. "Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights." *Applied Ergonomics*. 42(2), pp. 348–357. 2011. https://doi.org/10.1016/j.apergo.2010.08.005

[13] Charfuelan, M., Kruijff, G-J. "Analysis of speech under stress and cognitive load in USAR operations." In: *Natural Interaction with Robots, Knowbots and Smartphones*. (Mariani J., Rosset S., Garnier-Rizet M., Devillers L. (Eds)). pp. 275–281. Springer. 2014.

[14] Fant, G. "*Acoustic Theory of Speech Production*." The Hague, The Netherlands, Mouton, 1960.

[15] Wong, D. Y., Markel, J. D., Gray, J. A. H. "Least squares glottal inverse filtering from the acoustic speech wave form." *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 27(4), pp. 350–355. 1979. https://doi.org/10.1109/TASSP.1979.1163260

[16] Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne II, H. A., Hillman, R. E. "Mobile voice health monitoring using a wearable accelerometer sensor and a smart phone platform." *IEEE Transactions on Biomedical Engineering*. 59(11), pp. 3090–3096. 2012. https://doi.org/10.1109/TBME.2012.2207896

[17] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., Mehta, D. D. "Vocal and facial biomarkers of depression based on motor incoordination and timing." In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. ACM, 2014, pp. 65–72. https://doi.org/10.1145/2661806.2661809

[18] Plumpe, M. D., Quatieri, T. F., Reynolds, D. A. "Modeling of the glottal flow derivative wave form with application to speaker identification." *IEEE Transactions on Speech and Audio Processing*. 7(5), pp. 569–576. 1999. https://doi.org/10.1109/89.784109

[19] Gudnason, J., Brookes, M. "Voice Source cepstrum coefficients for speaker identification." In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 31 March-4 April 2008, pp. 4821-4824. https://doi.org/10.1109/ICASSP.2008.4518736

[20] Gudnason, J., Mehta, D. D., Quatieri, T. F. "Evaluation of speech inverse filtering techniquesusing a physiologically based synthesizer." In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, April 19-24, 2015, pp. 4245-4249. https://doi.org/10.1109/ICASSP.2015.7178771

[21] Le, P. N., Epps, J., Choi, E. H., Ambikairajah, E. "A study of voice source and vocal tract filter based features in cognitive load classification." In: 2010 20th International Conference on Pattern Recognition, Istanbul, Aug. 23-26, 2010, pp. 4516-4519. https://doi.org/10.1109/ICPR.2010.1097

[22] Yap, J. F., Epps, J., Ambikairajah, E., Choi, E. H. C. "Voice source features for cognitive load classification." In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, May 22-27, 2011, pp. 5700-5703. https://doi.org/10.1109/ICASSP.2011.5947654

[23] Gudnason, J., Thomas, M. R. P., Ellis, D. P. W., Naylor, P. A. "Data-driven voice source analysis and synthesis." *Speech Communication*. 54(2), pp. 199–211. 2012. https://doi.org/10.1016/j.specom.2011.08.003

[24] Patel, S., Scherer, K. R., Björkner, E., Sundberg, J. "Mapping emotions in to acoustic space: the role of voice production." *Biological Psychology*. 87(1), pp. 93–98. 2011. https://doi.org/10.1016/j.biopsycho.2011.02.010

[25] Heman-Ackah, Y. D., Sataloff, R. T., Laureyns, G., Lurie, D., Michael, D. D., Heuer, R., Rubin, A., Eller, R., Chandran, S., Abaza, M., Lyons, K., Divi, V., Lott, J., Johnson, J., Hillenbrand, J. "Quantifying the cepstral peak prominence, a measure of dysphonia." *Journal of Voice*. 28(6), pp. 783–788. 2014. https://doi.org/10.1016/j.jvoice.2014.05.005

[26] Fraile, R., Godino-Llorente, J. I. "Cepstral peak prominence: A comprehensive analysis." *Biomedical Signal Processing and Control*. 14, pp. 42–54. 2014. https://doi.org/10.1016/j.bspc.2014.07.001

[27] Alku, P. "Glottal wave analysis with pitch synchronous iterative adaptive filtering." *Speech Communication*. 11(2-3), pp. 109–118. 1992. https://doi.org/10.1016/0167-6393(92)90005-R

[28] Alku, P., Backstrom, T. "Normalized amplitude quotient for parametrization of the glottal flow." The *Journal of the Acoustical Society of America*. 112(2), pp. 701–710. 2002. https://doi.org/10.1121/1.1490365

[29] Waaramaa, T., Alku, P., Laukkanen, A. M. "The role of F3 in the vocal expression of emotions." *Logopedics, Phoniatrics, Vocology*. 3(4), pp. 153–156. 2006. https://doi.org/10.1080/14015430500456739

[30] Stuiver, A., De Waard, D., Brookhuis, K., Dijksterhuis, C., Lewis-Evans, B., Mulder, L. J. "Short-term cardiovascular responses to changing task demands." *International Journal of Psychophysiology*. 85(2), pp. 153–160. 2012. https://doi.org/10.1016/j.ijpsycho.2012.06.003